Final Report for Period: 03/2006 - 02/2007 **Submitted on:** 05/18/2007
Principal Investigator: Wills, Linda M. **Award ID:** 0092552
Organization: GA Tech Res Corp - GIT
Title:
CAREER: Automated Software Understanding for Retargeting Embedded Image Processing Software for Data Parallel Execution

Project Participants

Senior Personnel

Name: Wills, Linda
Worked for more than 160 Hours: Yes
Contribution to Project:
Linda Wills served as Principal Investigator on this project.

Post-doc

Graduate Student

Name: Sander, Samuel
Worked for more than 160 Hours: Yes
Contribution to Project:
Sam Sander performed doctoral research in developing a retargeting compiler for data parallel architectures for embedded multimedia applications. The focus is on handling variations in the pixel-per-processing element (PPE) ratio: the amount of image data directly mapped to each processing element, which is a key design parameter of SIMD architectures for image processing applications.

Name: Baumstark, Lewis
Worked for more than 160 Hours: Yes
Contribution to Project:
Lewis Baumstark performed doctoral research in retargeting sequential image processing programs to data parallel mechanisms (SIMD architectures and subword parallelism). He also developed new parallelism estimation techniques that help focus retargeting on sections of code that have the highest potential to yield a large speedup.

Name: Ryu, Soojung
Worked for more than 160 Hours: Yes
Contribution to Project:
Soojung Ryu's doctoral research involved developing automated methods for efficiently exploring the design space of memory configurations in embedded multimedia systems. Her techniques are a novel combination of cost models for evaluating area and energy efficiency, application retargeting techniques (at the assembly language level), and memory optimization transformations.

Name: Shah, Nidhi
Worked for more than 160 Hours: Yes
Contribution to Project:
Nidhi Kejriwal Shah explored the feasibility of performing runtime recognition and translation with architectural support. As part of her research, she investigated pattern detection mechanisms that monitor instruction caching structures (e.g., trace cache or fragment cache) to support runtime transformations.

Name: Bunchua, Santithorn

Worked for more than 160 Hours: Yes

Contribution to Project:

Santithorn Bunchua performed doctoral research on distributed register file architectures. He developed retargeting techniques to transform assembly-level code written in the traditional central register file style to an explicitly distributed register file architecture. The code retargeting framework focuses on minimizing the addition of register copying operations in the data routing stage and on spilling registers to other local register files when spare capacities are available.

Name: Melton, Roy

Worked for more than 160 Hours: Yes

Contribution to Project:

Roy Melton's doctoral research focused on parallelizing the spectral transform method used in climate and weather modeling. His parallelization targets scalable, massively data parallel implementations. He developed a tool (STOPPE) to predict parallel performance and a parallel spectral modeling decision support framework to efficiently explore configurations and evaluate their potential parallel performance.

Name: Kim, Hongkyu

Worked for more than 160 Hours: Yes

Contribution to Project:

Hongkyu Kim's doctoral research explored ways of exploiting the regular operand distribution patterns in multimedia applications to speed up operand movement within a datapath. He developed a new dynamic optimization mechanism (called dynamic SIMDization) and new low-latency operand transport techniques that dynamically cluster chains of dependent instructions at runtime. These dynamic techniques convert global communication (through expensive global broadcast networks) into local communication among functional units that are connected by a local dedicated network. These new operand transport mechanisms were designed based on empirical analysis of operand characteristics in a large body of multimedia benchmark programs.

Name: Valentine, Brian

Worked for more than 160 Hours: Yes

Contribution to Project:

Brian Valentine is researching SIMDization of image processing algorithms for video surveillance applications. His primary target is on high-performance embedded multimedia computing systems for real-time video surveillance.

Name: Huang, Tsai

Worked for more than 160 Hours: Yes

Contribution to Project:

Tsai Chi Huang focused his doctoral research on analyzing parallelism in network protocol processing applications. He developed an analytical method for estimating superscalar architecture performance using statistical information from communications protocol applications. This method predicts lower and upper bounds on the inherent instruction-level parallelism that can be exploited in these applications.

Undergraduate Student

Technician, Programmer

Other Participant

Research Experience for Undergraduates

Organizational Partners

Other Collaborators or Contacts

Activities and Findings

Research and Education Activities:

New compact, low-power implementation technologies for processors and imaging arrays can enable a new generation of portable video products. However software compatibility with large bodies of existing applications written in C prevents more efficient, higher performance data parallel architectures from being used in these embedded products. If this software could be automatically retargeted for explicitly data parallel execution, product designers could incorporate these architectures into embedded products. The key challenge is exposing the parallelism that is inherent in these applications but that is obscured by artifacts imposed by sequential programming languages.

This project developed reengineering techniques for retargeting embedded image and video processing software to data parallel execution. There is an enormous body of applications that implement a wide range of early image and video processing operations (e.g., convolution, wavelet decomposition, and motion estimation). These algorithms contain a significant amount of inherent parallelism since they often map operations across all pixels (or regions of pixels) in large two-dimensional image arrays. With the advent of data parallel multimedia extensions to instruction sets (e.g., MMX) and new massively parallel SIMD architectures, the potential to efficiently exploit this parallelism is emerging. However, most of these applications are written in C, which has no mechanism for specifying explicit parallelism, so the algorithms are implemented using nested loops. While modern compilers are capable of exposing limited loop-level parallelism, they target instruction level parallelism (ILP) using pipelining or superscalar execution. This is less efficient for these applications than data parallel execution which can provide a greater speedup at lower cost in area and energy.
The goal of this project is to automatically expose the inherent data parallelism in these applications to fuel a new generation of high performance, high efficiency embedded processing systems. Software understanding techniques have been developed to detect both the data parallel operations and the underlying data structures hidden in the code. This provides a deeper model of what computation is being performed so that a broader set of parallelization transformations can be accomplished and more opportunities for parallelization can be identified.

The specific activities conducted in this project are as follows.

Data parallelism extraction. Image processing algorithms are often inherently data-parallel, but the artifacts imposed by the sequential programming language (e.g., loops, pointer variables, linear address spaces) can obscure the parallelism and prohibit generation of efficient parallel code. We have developed a recognition-based approach for automatically extracting a data parallel program model from sequential image processing code and retargeting it to data parallel execution mechanisms. The explicitly parallel model presented, called multi-dimensional data flow (MDDF), captures a model of how operations on data regions (e.g., rows, columns, and tiled blocks) are composed and interact. The MDDF representation is based on an extension of the multi-dimensional synchronous dataflow (MDSDF) model of computation (originally developed for UC Berkeley's Ptolemy Classic). To extract an MDDF model, a partial recognition technique is used that focuses on identifying array access patterns in loops, transforming only those program elements that hinder parallelization, while leaving the core algorithmic computations intact. This allows the technique to be more generally applicable than approaches that require a specific core algorithm to be recognized. This automated retargeting technique has been implemented in a system, called PARRET (PARallel RETargeter).

Unlike traditional vectorization techniques, our approach recognizes patterns of data access (e.g., 2D element-wise multiplication). This deeper understanding of data movement and access patterns is more flexible in allowing multiple parallelization strategies, not restricted to just

vectorization. The technique has been applied to production image processing programs which are part of the Texas Instruments IMGLIB suite of applications for the TI TMS320C62xx line of DSPs.

Parallelism Estimation. We developed and validated a new technique to estimate inherent parallelism in sequential code. Retargeting existing code assets to parallel hardware execution mechanisms is difficult, but can provide significant increases in efficiency. It is desirable to estimate potential parallelism before undertaking the expensive process of reverse engineering and retargeting. We developed a new lightweight process for selection of likely loop candidates for retargeting. The technique takes a hybrid approach:
1) dependence data is collected by automatically annotating a program with instrumentation code and executing the annotated program and
2) static loop nesting information is collected and combined with the dependence data to produce a parallelism estimate. Together, the nesting model and dependence-distance profile are used to rank program loops with respect to its applicability for parallel retargeting.

Retargetable data parallel compilation. The rapid development of embedded multimedia products will lead to a diverse range of data parallel processing platforms. High-computational demand applications (e.g., real-time mpeg encoding) will drive specialized highly data parallel processors. Low-computational demand applications will continue to exploit current subword parallel instruction set extensions. The goal of this research activity is to develop a compiler that can translate a single application specification to any processor granularity along this diverse architectural spectrum. The particular challenge is to allow algorithms to be specified without architecture-dependent commitments to grain size.

In image processing applications, the grain size of the processing elements determines the number of pixels that are mapped to each processing element, which is called the pixel-per-processing-element ratio (PPE). The PPE determines how a pixel is accessed relative to another pixel. Varying the PPE ratio changes how a particular pixel location is determined relative to other pixels and how its value is retrieved (e.g., from local memory or via communication from other processors). When commitments to grain size (specifically, PPE ratio) are built into the specification of image processing algorithms, changing the granularity requires programs to be laboriously rewritten, which limits their portability and reusability. In this activity we developed compiler techniques that enable algorithm specifications to be written that are independent of grain size. The compilation techniques take a single high level source and (1) target it for data parallel execution on a wide range of processor granularities from a single processor to a massively parallel array, including (2) subword-level parallelism at each level of granularity as well as the (3) the ability to integrate random access expression into the specification.

We have integrated this compiler with the data parallel extraction technique (as the synthesis engine of PARRET). This is allowing us to retarget directly from C code to a broad range of data parallel platforms of varying processor granularities and subword sizes.

Dynamic SIMDization: Many multimedia applications have an abundance of inherent data-level parallelism (DLP) as well as ILP. However, modern general-purpose and embedded processors provide parallel execution mechanisms typically targeting only ILP, such as superscalar execution, and they are nearing the limit of what can be achieved with them. Additionally, increasing the number of functional units (FUs) to take advantage of existing ILP introduces a wire-dominated operand transport network that employs poorly scaling broadcast buses to distribute operands. Today, interconnect is becoming the limiting resource in integrated circuit fabrication and architectural focus is shifting from operand computations to operand communications.
We explored ways of exploiting the regular operand distribution patterns in multimedia applications to speed up operand movement within a datapath. We developed an execution mechanism that recognizes regular operand transport patterns and optimizes the operand movement in the dynamic execution environment. This exploits the data parallelism without increasing the communication overhead associated with operand movement within a datapath, especially for multimedia applications where the movement is highly regular.

This dynamic SIMDization work has two primary contributions:
1. It characterizes operand movement in media workloads from the perspective of support required for SIMD processing. Our study focuses on recognition of data access patterns between dependent instructions. We showed that data-parallel operation can be achieved by detecting and predicting stride values based on the recognized operand movement patterns.

2. It optimizes operand traffic by reducing needless communication within the datapath based on the operand characteristics. We focus on the 'intermediate' operands which are produced and only consumed within the same iteration of the loop or within the next iteration of the loop. Our empirical analysis reveals a large number of such short-lived, transient operands within the innermost loops of most multimedia applications and their communication can be localized. The major focus of this mechanism is on dynamic optimization to complement the large body of previous research requiring new ISAs and/or compiler support. In particular, we developed a new dynamic optimization mechanism (called dynamic SIMDization) and new low-latency operand transport techniques that dynamically cluster chains of dependent instructions at runtime.

These dynamic techniques convert global communication (through expensive global broadcast networks) into local communication among functional units that are connected by a local dedicated network. These new operand transport mechanisms were designed based on empirical

analysis of operand characteristics in a large body of multimedia benchmark programs.

Distributed register file retargeting. Traditional processors are designed with a single central register file which has several drawbacks in terms of required die area, access time, and energy consumption. High levels of parallelism dictate a large register file with a large number of registers and a large number of access ports per register. A distributed register file architecture can be used to alleviate this problem for new parallel architectures. We developed a code retargeting framework to transform assembly-level code written in the traditional central register file style to an explicitly distributed register file architecture. The code retargeting framework focuses on minimizing the addition of register copying operations in the data routing stage. Moreover, a register allocation technique for distributed register files is being developed with the ability to spill registers to other local register files when spare capacities are available. This not only improves performance by using a fast spilling mechanism (rather than directly spilling to memory every time) but also results in efficient use of register file resources. Scheduling of data routing operations concurrent with functional unit operations is examined as a mechanism to reduce the penalty incurred by extra register copy cycles. The effectiveness of the code retargeting scheme was assessed by applying it to Mediabench and Spec 2000 benchmarks.

Parallelization of spectral transform method. This research addresses the realizable parallel performance of the spectral transform method used in climate and weather models, especially in the context of scalable, massively data parallel implementations. In contrast to existing analyses, which are abstract and simplistic, this research provides comprehensive, quantitative computational characterizations of the spectral transform, its parallelism, and its parallel performance for the general case. The analysis results in a decision support framework for parallel spectral modeling to configure both hardware and model for optimal forecast speed and accuracy from available resources.

Parallelism in communications protocols: Increasing diversity in telecommunication workloads leads to greater complexity in communication protocols. This occurs as channel bandwidth rapidly increases. These factors result in larger computational loads for network processors that are increasingly turning to high performance microprocessor designs. This research developed an analytical method to estimate the performance of instruction level parallel (ILP) processors executing network protocol processing applications. Statistical instruction dependency information extracted while executing an application is used to predict upper and lower bounds for throughput. The analytical method is much less expensive than cycle-accurate simulation, but reveals similar throughput performance predictions.

Efficient storage usage in embedded SIMD systems. A key bottleneck in realizing the data parallelism inherent in multimedia applications is memory usage and efficient data movement. Storage is expensive in terms of area, energy, and latency û costs which are acutely felt in portable embedded multimedia systems. The objective of this research is to achieve greater area-efficiency and energy-efficiency for on-chip memory. The approach is to develop methods to efficiently explore on-chip storage configurations that are optimal for a given set of applications. The methods are a novel combination of cost models for evaluating area- and energy-efficiency, application retargeting techniques (at the assembly language level), and memory optimization transformations. We have developed an automated retargeting technique that is used in analyzing the storage requirements of a program over a range of storage (memory and register file) configurations. This type of retargeting is too expensive and labor intensive to perform manually during exploration. This is particularly true for hand-coded assembly language programs that are optimized for specific embedded memory designs. The memory optimization transformations we have developed synergistically optimize both code (register and memory references) and the memory hierarchy design parameters, guided by energy and area cost models.

Curriculum development: We developed a new required core undergraduate course, which has been offered every semester since Fall 2004, on 'Mechanisms for Computing Systems.' We developed MiSaSiM, a program execution simulator that supports flexible exploration of an assembly program's execution. It enhances program understanding by providing reversible trace navigation and automated recognition features. This educational tool uses some of the underlying analysis technology being developed in this project.

We also developed a laboratory project for an advanced computer architecture graduate course which explores data parallel programming for image processing applications. The goal is to give the students practice in understanding data parallel programming and in converting sequential concepts to parallel algorithms. An architectural SIMD simulator is used to allow students to test and debug their programs. The project also includes an exercise in contrasting estimates of performance for instruction-level versus data-level parallel versions of the code written by the students.

Finally, we recently developed a new graduate course on embedded video surveillance systems, which was offered for the first time in Spring 2007. This hands-on course brings together material from many fields, such as digital signal processing, computer architecuture, microelectronics, computer vision, and parallelization (including the research conducted under this grant).

Findings:

We produced a number of interesting results over the course of this project.

Data parallelism extraction. We have automated our recognition-based retargeting technique in a system, called PARRET (PARallel RETargeter). We have used it to successfully parallelize a set of production programs from the Texas Instruments IMGLIB suite of applications for the TI TMS320C62xx line of DSPs. Our MDDF representation is architecture-independent, giving it the flexibility to target multiple architecture types. To demonstrate the generality of our approach we retargeted the extracted MDDF model of each benchmark program to two different data parallel targets:

1. The first target is subword parallel instructions found in multimedia instruction set extensions, in particular, Intel's SSE2: Compiling for multimedia instruction set extensions is constrained by the unit stride array accesses required for packed vector data. For example, vector data types of mixed precision within a loop body create non-unit stride with respect to each other, complicating the parallelization process as corresponding array elements are unaligned in the packed vectors. PARRET's high-level multi-dimensional representation facilitates properly aligning these mixed vector data types by exposing the appropriate place in the algorithm to generate data type promotion or demotion code. Speedups of 2 to 27 times over baseline sequential execution are shown for the retargeted loop kernels. A comparison with a commercial vectorizing compiler, the Intel C/C++ compiler (ICL) demonstrated PARRET's ability to parallelize a wider range of loop-based applications than traditional vectorization; there were seven benchmarks that ICL could not vectorize, but PARRET was able to parallelize. These results were published in the 3rd International Workshop on Embedded Computing, held in conjunction with the 2006 International Conference on Parallel Processing (ICPP-06) in August, 2006.

2. The second target is a massively parallel SIMD processor array, in particular, a representative machine, SIMPil: The retargeted applications yield a potential execution throughput limited only by the number of processing elements, exceeding thousands of instructions per cycle in large-scale SIMD implementations. We showed that the retargeted SIMPil code is able to execute at a much higher instructions-per-cycle (IPC) rate than the original code running on a sequential, superscalar baseline û between two and four orders of magnitude greater performance than the typical sequential execution. These results have been published in the paper 'Retargeting Sequential Image-Processing Programs for Data parallel Execution' which appeared in IEEE Transactions on Software Engineering, February 2005.

Sequential Distortions. An unexpected benefit of the MDDF representation is that it enables us to easily detect and remove (if desired) 'sequential distortions' that arise in image processing applications. Objectives of image processing applications often differ from traditional programming domains. For example, image and video processing implementations often alter the strict algorithm specification, introducing slight distortions in the results that are imperceptible to human senses (e.g., widely-used impulse noise filters introduce slight blurring to noise-free regions that cannot be detected by the human eye). In some cases, sequential implementations of algorithms do not accurately express the theoretical function (e.g., pure convolution) because doing so would require the overhead of checking complex boundary conditions. For example, to simplify and speed up the implementation, pixel neighborhoods may be allowed to 'wrap around' between opposites edges of an image since they are adjacent in linear memory. This introduces imperceptible errors at the boundaries of the image. In retargeting these applications to data parallel execution, it may not be desirable to preserve these optimizations and the human reengineer needs to decide whether they should be retained. Unlike typical reengineering tasks, in this domain, the original program might not contain an accurate specification of the intended algorithm, but instead might contain approximations or optimizations that only make sense in the sequential implementation. This necessitates augmenting the reengineering process by detecting these sequential distortions, estimating their effect, and confirming with the human reengineer that the sequential distortions should be removed when retargeting to the parallel version. MDDF exposes the sequential distortions resulting from sequential optimizations and approximations. Our recognition process estimates the extent of these distortions and, with the approval of the human reengineer, automatically retargets to an improved parallel version without the distortions. These results appeared in the paper mentioned above that was published in IEEE Transactions on Software Engineering, February 2005.

Parallelism Estimation. We developed a technique for estimating the amount of data-level parallelism available in loop-based algorithms. The technique was tested on a suite of programs from the TI IMGLIB suite and the MediaBench suite. Correctness was demonstrated by comparing with known expected results. Timing measurements were performed, with the estimation of most benchmarks completing in under ten minutes, and all completing within an hour. The results show the correctness of this method, its increased performance ù on average, 334 times faster over earlier simulation-based techniques û and that it scales linearly with input program size.

Retargetable data parallel compilation. To show that the same high-level source can be retargeted to widely different architectures with varying granularity, performance comparisons were made from simulation results using both a SIMD array target architecture and a general-purpose processor target architecture. These validation experiments yield three main results. First, they demonstrate the ability to retarget the same PPE-independent specification to processors of varying grain sizes û ranging from sequential processors (PPE equal to image size) to massively data-parallel multiprocessor systems (PPE equal to one) û with little loss of execution time efficiency or quality of code generated, compared with PPE-dependent source programs. Second, for single and multiprocessor targets with SIMD instruction set extensions, the research compiler code produced speedups linear with subword count on targets with various word widths (for configurations with a PPE ratio relatively larger than the wide-word size). Third, the use of hybrid algorithms provided comparable performance to applications written using architecture-dependent source code, which would not be otherwise possible using architecture-independent source code. In addition to speedup

from hybrid algorithm usage, concurrently targeting subword-level parallelism yielded linear speedups for configurations with a PPE ratio relatively larger than the wide-word size. A paper based on this work 'The impact of grain size on the efficiency of embedded SIMD image processing architectures' appeared in the Journal of Parallel and Distributed Computing in November 2004. The paper quantitatively evaluates the impact of PPE ratio on system performance and efficiency for focal-plane SIMD image processing architectures by comparing throughput, area efficiency, and energy efficiency for a range of common application kernels using architectural and workload simulation. Using these evaluation techniques (application grain size retargeting combined with area and energy technology modeling), a new class of efficient, embedded SIMD architectures for image processing can be designed.

Dynamic SIMDization: The performance of modern ILP processors is approaching their limits due to the limited instruction parallelism and complex operand communication mechanism needed to enhance parallelism. By analyzing the characteristics of the operand transport during the execution, we have shown that inherent data parallelism can be detected dynamically by exploiting regular operand transport patterns. Our dynamic mechanism can also control the operand traffic based on the dependence between instructions and loop iterations, resulting in minimizing the latency associated with operand movement. In particular, we first introduced the idea of instruction clustering which groups dependent instructions for the innermost loop. Second, we characterized the operands connecting the instructions and loop iterations. Third, we presented a dynamic SIMD mechanism which separates the data parallel and non-parallel region from the stride prediction scheme, schedules them to the dedicated SIMD unit, and bypasses the results of instructions through the dedicated paths. Our results show that the overall performance gains over the conventional ILP processors are 87% for 8-way and 114% for 16-way on average. When wire delay latency is factored in, the performance gap increases to 109% for 8-way and 159% for 16-way respectively. Most of the speedup comes from exploiting more parallelism and localizing most operand communication. The dynamic SIMDization technique and results have been published in a paper presented at the IEEE International Workshop on Multimedia Signal Processing (MMSP06) in October 2006. The instruction clustering technique for multimedia applications was published in the IEEE International Symposium on Multimedia (ISM 2005) in December 2005.

Distributed register file retargeting. Operand transport is a critical problem in traditional high-performance processor architectures that use central register files and bypass networks. We are exploring retargeting to a fully distributed register file organization to reduce register access port demands, signaling distance, and implementation complexity. The main challenge in retargeting is that register file coherency must be maintained by inserting explicit register transfer operations to minimize the IPC penalty. Eager and multicast transfer techniques are explored to reduce the IPC penalty for static and dynamic execution. The average IPC penalties are 22% and 23% for the dynamic and static approaches, respectively, in an 8-way configuration. These penalties are offset by a 41% reduction in operand access time, 86% reduction in area, and 87% reduction in energy consumption of a distributed register file compared to a central register file structure. Overall performance speedup is achieved through significant improvement in processor clock frequency. A journal paper describing these results is currently under review. We also published results in the International Conference on Computer Design (ICCD 2003) that we obtained from running retargeted SpecInt 2000 code on a simulated distributed register file (DRF) architecture, using Simplescalar extended with DRF and register transfer capabilities.

Parallelization of spectral transform method. The products of this activity are the quantitative computational characterization of the spectral transform, its parallelism, and its parallel performance. Existing computational expressions for the spectral transform are abstract and simplistic (because they account for only the highest order terms); this comprehensive, detailed computational analysis of the spectral transform yields closed-form expressions for all of its computational components. Similarly, in contrast to existing parallel spectral transform analyses, this research not only identifies sources of parallelism within the transform, but also derives closed-form expressions for all parallel transform computations and communication and thus identifies the degree and type of parallelism within the transform. This research culminates in the development of the Spectral Transform Operation and Parallel Performance Estimator (STOPPE), which predicts the parallel performance of the spectral transform for any processor configuration or model data resolution. In comparison with measured performance results of a spectral model, STOPPE estimates reflect observed performance trends, and its predicted speedups for various parallel configurations fall within 3.5% of actual measured speedups. Thus, this research serves as a guide for efficient implementation of current models as well as development of future models on massively parallel computers.

Parallelism in communications protocols: an analytical method was developed to estimate superscalar architecture performance (measured in instructions per cycle, IPC) on network protocol processing applications. Results using UDP/TCP/IP applications show the simulated IPC values fall between the analytically derived upper and lower bounds, validating the model. The analytical method is much less expensive than cycle-accurate simulation, but reveals similar throughput performance predictions. This allows the architectural design space for network superscalar processors to be explored more rapidly and comprehensively, to reveal the maximum IPC that is possible for a given application workload and the available hardware resources. This research was published in the journal Performance Evaluation in October 2006.

Efficient storage usage in embedded SIMD systems. An analysis method for assessing storage needs and costs of a given application automatically retargeted across a spectrum of storage configuration designs was developed. Energy and area models have been developed for use in exploring the design space of on-chip memory design configurations. This technique has been applied to both embedded SIMD and MIPS applications (including Vector Quantization, median filtering, convolution, image rotation, and TAK). The technique has been effective in finding the optimal storage configuration in terms of area and energy efficiency. In particular, this technique found that a SIMD processing

element achieves optimal area and energy efficiency with a register file containing between 8 and 12 words for a given workload. This configuration is between 15% and 25% more area and energy efficient than other memory configurations being considered.

Curriculum development: The educational tool MiSaSiM developed using program analysis technology underlying this project is an integral part of a systems-oriented computer engineering curriculum at Georgia Tech. It provides students with an opportunity to assess the functionality, performance, efficiency, and cost of a program. Novel features include reversible trace navigation, visualizations, and support for distance education delivery. This work is designed to better prepare computer engineering students for system-oriented careers. It introduces them to important design considerations beyond functionality such as, performance, efficiency, reliability, and cost. A paper on this course and the MiSaSiM educational tool will appear in the Frontiers in Education (FIE) conference in October 2007.

In addition, the SIMD laboratory project in the graduate course ECE6100 has generated enthusiastic discussions in class and outside of class. Student project reports were generally insightful and demonstrated an understanding of data parallel programming issues.

Training and Development:

This project has provide graduate research experience to nine graduate students. Seven have completed their doctorates and one earned a masters degree. The other students is currently in the second year of his doctoral program. Four of these graduates are in academic faculty positions and three have research positions at industrial research laboratories.

Outreach Activities:

Throughout this project, Dr. Wills and her graduate students have been engaged in a variety of outreach activities to attract high school students and undergraduates to engineering. They have specifically reached out to students from underrepresented groups through the following programs:

1. Partnering in Transitioning to Tech (PITT): this program, directed by Dr. Linda Wills, is aimed at increasing participation by women and underrepresented minority students in engineering. It focuses specifically on students from underrepresented groups who are following nontraditional paths to an engineering degree (e.g., through a dual degree program or transferring in from a community college). Recent reports by AAAS and NAE/NRC reveal the critical role of ommunity colleges and nontraditional educational paths in information technology, computing, and engineering education, particularly in opening career opportunities for students from underrepresented groups. The PITT program includes student outreach and preparation, peer partnerships, mentoring, tutoring, faculty advising, and undergraduate research experiences. The PITT program has been extremely successful. The overall retention rate for all PITT students from Fall 2004-Fall 2006 is 97% (28/29 students), in contrast with a previous 58% retention rate in Spring 2004. The average cumulative GPA of 2.65 is an improvement over the Spring 2004 average cumulative GPA of 2.03. This program began as a pilot program in the School of Electrical and Computer Engineering (ECE) for dual-degree students who transfer from historically black Atlanta University Center Colleges. As a result of the success of this program in ECE, PITT was expanded in 2006 to Biomedical Engineering (BME) and Mechanical Engineering (ME) with seed funding from Lockheed-Martin. We are pursuing additional funding as well as corporate sponsorship to sustain and grow this program to include all underrepresented minority tranfer students in all engineering disciplines in the College of Engineering.

2. Summer Undergraduate Research in Engineering/Science (SURE): Dr. Wills and her graduate students engaged undergraduate minority students in research related to this grant as faculty mentors and graduate student mentors in the SURE program sponsored by NSF (REU). Through close interaction with the students on this research, the mentors try to provide appropriate role models and encourage the students to pursue graduate degrees.

3. Opportunity Scholars Program is a program specifically designed to increase retention and success of first- and second-year underrepresented computer students by pairing them with graduate student mentors and their advisors, who involve them in their research during the academic year. Dr. Wills and two of her graduate students (Nidhi Kejriwal Shah, who was supported on this project, and Cameron Craddock, who was supported on another NSF project) mentored four undergraduates (two African-American and two female students) as part of this program.

4. Engineering and Computing Career Conference (ECC): Dr. Wills participated in the ECC Conference (Fall 1999-2004) for attracting high school students to engineering in which she discussed this

research as part of her presentation. She also organized student and faculty panels. Most of the participants were women and underrepresented minorities.

The research opportunities we provided to students through this grant and through the programs listed above helped to promote diversity in our field.

Journal Publications

Linda Wills, Tarek Taha, Lewis Baumstark, Scott Wills, "Estimating Potential Parallelism for Platform Retargeting", Working Conference on Reverse Engineering (WCRE), p. 55, vol. , (2002). Published

Lewis Baumstark and Linda Wills, "Exposing Data-Level Parallelism in Sequential Image Processing Algorithms", Working Conference on Reverse Engineering (WCRE), p. 245, vol. , (2002). Published

Randall Janka, Linda Wills, and Lewis Baumstark, "Virtual Benchmarking and Model Continuity in Prototyping Embedded Multiprocessor Signal Processing Systems", IEEE Transactions on Software Engineering, p. 832, vol. 28, (2002). Published

Santithorn Bunchua, Scott Wills, Linda Wills, "Reducing Operand Transport Complexity of Superscalar Processors using Distributed Register Files", Proceedings of the International Conference on Computer Design (ICCD), p. 532, vol. , (2003). Published

Lewis Baumstark, Murat Guler, and Linda Wills, "Extracting an Explicitly Data-Parallel Representation of Image-Processing Programs", Proceedings of the 10th Working Conference on Reverse Engineering (WCRE), p. 24, vol. , (2003). Published

Antonio Gentile, Sam Sander, Linda Wills and Scott Wills, "The impact of grain size on the efficiency of embedded SIMD image processing architectures", Journal of Parallel and Distributed Computing, p. 1318, vol. 64, (2004). Published

Lewis Baumstark and Linda Wills, "Retargeting Sequential Image-Processing Programs for Data parallel Execution", Transactions on Software Engineering, p. 116, vol. 31, (2005). Published

Tsai Chi Huang, Linda Wills, Roy Melton, and Cecil Alford, "Predicting Communication Protocol Performance on Superscalar Architectures using Instruction Dependency", Performance Evaluation, p. 939, vol. 63, (2006). Published

Samuel Sander and Linda Wills, "Retargeting Image-Processing Algorithms to Varying Processor Grain Sizes", Journal of Embedded Computing, p. , vol. , ( ). Invited submission due 15 June 2007

Roy Melton and Linda Wills, "An Analysis of the Spectral Transform Operations in Climate and Weather Models", SIAM Journal on Scientific Computing, p. , vol. , ( ). Submitted

Lewis Baumstark and Linda Wills, "Multidimensional Dataflow-based Parallelization for Multimedia Instruction Set Extensions", Proc. of the 3rd Int. Workshop on Embedded Computing, held in conjunction with the 2006 International Conference on Parallel Processing (ICPP-06), p. 1, vol. , (2006). Published

Hongkyu Kim, Scott Wills, and Linda Wills, "Reducing Operand Communication Overhead using Instruction Clustering for Multimedia Applications", Proceedings of the IEEE International Symposium on Multimedia (ISM 2005), p. 345, vol. , (2005). Published

Hongkyu Kim, Scott Wills, and Linda Wills, "Optimizing Operand Transport using Dynamic SIMDization in Multimedia Systems", Proc. of the IEEE International Workshop on Multimedia Signal Processing, p. 372, vol. , (2006). Published

Linda Wills and Scott Wills, "MiSaSim: A Resource-Aware Programming Environment for Computer Systems Engineering Education", Frontiers in Education Conference (FIE07), p. 1, vol. , (2007). Published

Books or Other One-time Publications

Web/Internet Site

Other Specific Products

Contributions

Contributions within Discipline:

Scientific, symbolic, and multimedia applications present diverse computing workloads with different types of inherent parallelism. Tomorrow's processors will employ varying combinations of parallel execution mechanisms to efficiently harness this parallelism. The explosion of consumer products that incorporate high performance embedded computing will increase the stratification of the processor design space. However, existing code assets are limited to sequential expression of what should be highly parallel algorithms. Retargeting to parallel mechanisms is difficult, but can provide significant increases in efficiency. This research is contributing techniques
(1) for estimating potential parallelism before undertaking the expensive process of reverse engineering and retargeting and
(2) for automatically extracting data level parallelism from sequential image processing programs.

This contributes algorithms and program representations that go beyond those developed for traditional parallelization and vectorization of code. In particular, the reengineering techniques preserve the spatial and temporal adjacency of image data. This allows parallelism to be exploited efficiently, not just in performance, but also in area-efficiency and energy-efficiency for compact, low-cost, long battery-life portable video systems.

Contributions to Other Disciplines:

In addition to extracting the inherent prarllelism in multimedia workloads, this research also explored the potential for parallelism in climate and weather modeling. As part of an analysis of scientific workloads, our focus was drawn to the spectral transform, which is a method commonly used in climate and weather models but which is known to be a key bottleneck of spectral model execution, in msome cases responsible for 70% to 90% of a model's serial execution time. This research provides comprehensive, quantitative computational characterizations of the spectral transform, its inherent parallelism, and its parallel performance for the general case. The analysis results in a decision support framework for parallel spectral modeling to configure both hardware and model for optimal forecast speed and accuracy from available resources.

Our outreach activities also contribute to other disciplines beyond electrical and computer engineering (ECE). The Partners in Transitioning to Tech (PITT) program, which Dr. Wills directs, aims to promote equal access to top-tier engineering education by increasing the enrollment, retention, and graduation rates of underrepresented transfer students pursuing engineering degrees. This program was started in ECE, but is transferring to other engineering disciplines (Mechanical Engineering and Biomedical Engineering, so far). Our plans are to expand it to all the engineering schools at Georgia Tech. A paper on this program has been published and presented at the Frontiers in Education Conference.

Contributions to Human Resource Development:

This project provided research opportunities for nine doctoral students in computer engineering. Two of the students are women and another student is African-American.

Also, this project was described as part of an Engineering and Computing Career Conference for high school students (mostly women and minorities) to provide exposure to research and computer engineering.

Dr. Wills and two of her graduate students served as mentors in the Summer Undergraduate Research in Engineering/Science (SURE) program and in the Opportunity Scholars Program for a total of six undergraduates, which involved engaging the students in this research.

Dr. Wills also established a new mentoring program (PITT) that targets undergraduates who transfer into Georgia Tech at the junior level as dual-degree students from historically black Atlanta University (AU) Center Colleges and Fort Valley State University. This group faces special challenges as they join the program mid-stream. This new program addresses the need among these students to have opportunities for full academic integration in the engineering program at Georgia Tech, including exposing the students to opportunities for undergraduate research within this project.

This project is also integrating tools and research problems into the educational curriculum of undergraduate and graduate students in computer engineering in three specific ways.

1. Dr. Linda Wills (with Dr. Scott Wills) developed a new core computer engineering undergraduate course 'Mechanisms for Computing Systems,' which shows how execution and storage mechanisms support high-level programming languages and operating systems. We developed a new educational tool, MiSaSiM, to support design projects in this course. MiSaSiM is a program execution simulator that supports flexible exploration of an assembly program's execution. It enhances program understanding by providing reversible trace navigation and automated recognition features (leveraging from expertise gained in this research project). A paper describing MiSaSiM and a variety of design projects developed for the course has been accepted to the Frontiers in Education (FIE07) conference.

2. A new laboratory project based on data parallel programming has been incorporated into the core graduate level advanced computer architecture course at Georgia Tech.

3. Dr. Linda Wills (with Dr. Scott Wills) developed a new graduate course on embedded video surveillance systems. This hands-on course brings together material from many fields, such as digital signal processing, computer architecuture, microelectronics, computer vision, and parallelization (including the research conducted under this grant).

Contributions to Resources for Research and Education:

Contributions Beyond Science and Engineering:

Categories for which nothing is reported:

Organizational Partners

Any Book

Any Web/Internet Site

Any Product

Contributions: To Any Resources for Research and Education

Contributions: To Any Beyond Science and Engineering